Australian Journal of Basic and Applied Sciences, 5(2): 157-164, 2011
ISSN 1991-8178

Approximation by Normal Distribution in Generalized Allocation Scheme and its Applications

SHERZAD MIRAKHMEDOV

GIK Institute of Engineering Sciences & Technology, Topi-23640. Swabi. Pakistan

Abstract: A random vector of frequencies of cells in the generalized allocation scheme of particles into cells is defined through conditional distribution of a random vector with integer, non-negative and independent components given their sum. A wide class of statistics, viz. the sum of functions of the frequencies is studied. The Berry-Esseen type bound is obtained. This result applied for a problem of testing of hypothesis on the probabilities of output outcomes in probabilistic set-up without memory.

Key words: Allocation scheme, binomial distribution, multinomial distribution, sample sum, probabilistic setup.

INTRODUCTION

Many combinatorial problems in probability and statistics can be formulated and best understood by using appropriate allocation schemes (alternatively known as urn models). Such models naturally arise in statistical mechanics, clinical trials, cryptography etc. The properties of several types of the random allocation scheme have been extensively studied in the probabilistic and statistical literature, see: Jonson, and Kotz (1977) and Kotz and Balakrishman (1997). The classical allocation scheme supposes an equiprobable allocation of *n* particles into a finite number of cells, say *N*, i.e. the probability of a particle falling into any particular cell is 1/N. There are several generalizations of the classical scheme. One of generalization is multinomial allocation scheme with finite number of cells. Here the probability of a particle falling into *m*-th cell is

$p_m > 0 p_1 + ... + p_N = 1$. Let η_m be a number of particles in the *m*-th cell after allocating of all *n* particles, then one can be easily checked that $\mathcal{L}(\eta_1,...,\eta_N) = \mathcal{L}(\xi_1,...,\xi_N \,|\, \xi_1 + ... + \xi_N = n)$ where $\xi_1,...,\xi_N$ are independent Poisson random variables with mean np_m respectively, *L (Y)* is the distribution of the random vector (r.vec.) Y. On the basis of this property we introduce the following generalized allocation scheme (GAS) of *n* particles into countable set of cells.

Let $\xi = (\xi_1, \xi_2 ...)$ be a r.vec. with independent, integer and non-negative components such that $P\{\xi_1 + \xi_2 + ... = n\} > 0$ $\eta = (\eta_1, \eta_2, ...)$ be a r.vec. the distribution of which defined as

$$\mathcal{L}(\eta_1, \eta_2, ...) = \mathcal{L}(\xi_1, \xi_2, ... | \xi_1 + \xi_2 + ... = n) \quad (1.1)$$

Note that (1.1) implies $P\{\eta_1 + \eta_2 + ... = n\} = 1$ The probabilistic model defined by (1.1) determines the GAS: *n* particles are allocated into countable number (finite or infinite) of cells; the distributions of the r.vec. through (1.1) defines the allocation scheme; here η_m is a number of particles in the *m*-th cell after allocating

Corresponding Author: SHERZAD MIRAKHMEDOV, GIK Institute of Engineering Sciences & Technology, Topi-23640. Swabi. Pakistan
E- mail: shmirakhmedov@yahoo.com

of all *n* particles. We are interesting in the following class of statistics, $R_n(\eta) = \sum_{m=1}^{\infty} f_{m,n}(\eta_m)$ (1.2)

where $f_{1,n}(x), f_{2,n}(x), ..$ are Borel functions of non-negative *x* such that the series (1.2) is convergence with probability equal to one. The following three examples of GAS and DS are most common in application.

Example 1:

Let $\mathcal{L}(\xi_m) = Bi(d_m, p)$ be binomial distribution with parameters $d_m > 0$ and , $p \in (0,1)$ and $P\{\xi_m = 0\} = 1$ for m > N, where integer N >1. Then, for any $p \in (0,1)$ the r.vec. η has a multidimensional hypergeometric distribution:

$$P\{\eta_1 = k_1, ..., \eta_N = k_N\} = \binom{D_N}{n}^{-1} \prod_{m=1}^{N} \binom{d_m}{k_m}$$

where $D_N = d_1 + ... + d_N$, $k_1 + ... + k_N = n$, $0 \le k_m \le d_m$ *m = 1,....., N*. This GAS corresponds to the sample scheme without replacement from the stratified finite population. A sample sum is type (1.2) statistic.

Example 2:

If $\mathcal{L}(\xi_m) = \Pi(zp_m)$ be a Poisson distribution with mean Zp_m , arbitrary Z > 0, and $p_m > 0$, m = 1,2 .....; $p_1 + p_2 + ;.. = 1$ then we deal with multinomial allocation scheme. The r.vec. η has the multinomial distribution M ($n, p_1, p_{2.....}$) with may be infinite number of outcomes; if $p_m > 0$ m = 1,....., N, and $p_m = 0$ for m > N , then we have above mentioned multinomial allocation scheme. The classical chi-square, likelihood ratio statistics and empty boxes statistic are examples of type (1.2) statistics. Another example is presented below in Sec 4.

Example 3:

Let $\mathcal{L}(\xi_m) = NBi(d_m, p)$ be negative binomial distribution with $d_m > 0$ and arbitrary p ε (0,1),

m = 1,.....,N and $P\{\xi_m = 0\} = 1$ for m > N , where integer N > 1. Then

$$P\{\eta_1 = k_1, ..., \eta_N = k_N\}, \quad (1.2)$$

$$= \binom{D_N + n - 1}{n}^{-1} \prod_{m=1}^{N} \binom{d_m + k_m - 1}{k_m}$$

where $D_N = d_1 + ...+ d_N$. Such specification of GAS corresponds to the multicolor Pólya- Egenberger urn model: consider an urn containing $\omega_m > 0$ balls of color m = 1,2....., N, a single ball is drawn from the urn, recorded and then returned together with $k \in$ balls of the same color, at each draw the probability to be drawn is the same for all balls. Assuming $-nk < \min(\omega_1,...,\omega_N)$ the drawing is repeated *n* times. Let r.v. η_m be a number of balls of *m*-th color appeared in the sample of size *n*. The r.vec. $\eta = (\eta_1...\eta_N)$ has distribution (1.2) with $d_m = \omega_m / k$. A number of colors that appeared in the sample exactly times is type (1.2) statistic.

For detailed information on GAS and its applications and references, consult, for example, Jonson and Kotz (1977), Kotz and Balakrishan (1997), Mirakhmedov (1996, 2007).

II. Bartlett Type Formula:

We assume that the series written below are convergent for each *n*. Put

$$A_n = \sum_{m=1}^{\infty} E\xi_m \qquad B_n^2 = \sum_{m=1}^{\infty} Var\xi_m$$

$$x_n = (n - A_n)/B_n \qquad \Lambda_n = \sum_{m=1}^{\infty} Ef_m(\xi_m)$$

$$\gamma_n = \frac{1}{B_n^2}\sum_{m=1}^{\infty} \operatorname{cov}\left(f_m(\xi_m), \xi_m\right)$$

$$g_m(y) = f_m(y) - Ef_m(\xi_m) - \gamma_n(y - E\xi_m)\,.$$

$$\sigma_n^2 = \sum_{m=1}^{\infty} Var\, g_m(\xi_m) = \sum_{m=1}^{\infty} Varf_m(\xi_m) - B_n^2\gamma_n^2$$

Under quite weak conditions, we have

$$ER_n(\eta) = \Lambda_n + x_n B_n \gamma_n$$

$$-\frac{1-x_n^2}{2B_n^2}\sum_{m=1}^{\infty} Eg_m(\xi_m)(\xi_m - E\xi_m)^2\left(1+o(1)\right) \qquad VarR_n(\eta) = \sigma_n^2\left(1+o(1)\right)$$

as $n \to \infty$ and $N = N(n) \to \infty$ Also

$$\hat{R}_n(\eta) = R_n(\eta) - \Lambda_n - x_n B_n \gamma_n = \sum_{m=1}^{\infty} g_m(\eta_m)$$

and

$$\sum_{m=1}^{\infty} Eg_m(\xi_m) = 0,\ \sum_{m=1}^{\infty} \operatorname{cov}\left(g_m(\xi_m), \xi_m\right) = 0.$$

For any measurable function φ such that $E\left|\phi(\xi_1, \xi_2, ..., \xi_N)\right| < \infty$ one has the following Bartlett's type formula:

$$E\phi(\eta_1,\eta_2,...) = \frac{1}{2\pi P\{\zeta_n = n\}} \int_{-\pi}^{\pi} E\phi(\xi_1,\xi_2,...)\exp\{i\tau(\zeta_n - n)\}\,d\tau$$

where $\zeta_n = \xi_1 + \xi_2 + ...$ Set

$$\hat{g}_m = g_m(\xi_m)/\sigma_n,\quad \hat{\xi}_m = (\xi_m - E\xi_m)/B_n \quad \psi_m(t,\tau) = E\exp\left\{it\hat{g}_m + i\tau\hat{\xi}_m\right\}$$

$$\Theta_n(t,x_n) = \frac{1}{\sqrt{2\pi}} \int_{-\pi B_n}^{\pi B_n} e^{-i\tau x_n} \prod_{m=1}^{\infty} \psi_m(t,\tau)\,d\tau.$$

This formula together with inversion formula for the local probability $P\{\zeta_n = n\}$ implies

$$\varphi_n(t,x_n) \overset{def}{=} Ee^{it\sigma_n^{-1}\hat{R}_n(\eta)} = \frac{\Theta_n(t,x_n)}{\Theta_n(0,x_n)} \tag{2.1}$$

It is important that in Bartlett type formula (2.1) the integrand in the definition of $\Theta_n(t,x_n)$ is the characteristic function of a sum of independent two-dimensional random vectors. This fact allows using the method characteristic function well developed for the sum of independent r.vec's.

Formula (2.1) can be extended for similar to type (1.2) statistics defined on several independent GASs, and asymptotic results can be derived also. One of specified statistic in such of GAS is considered by Mirakhmedov S.S. and Mirakhmedov S.M.(2009).

III. Main Result:

Let Φ (u) be the standard normal distribution function and $\mathrm{P}_n(u) = P\left\{\hat{R}_n(\eta) < u\sigma_n\right\}$Denote

$$\kappa_{j,n} = \sum_{m=1}^{\infty} E\left|\hat{\xi}_m\right|^j \qquad \beta_{j,n} = \sum_{m=1}^{\infty} E\left|\hat{g}_m\right|^j$$

$$M_n(T) = \inf_{T\le|\tau|\le\pi} \sum_{m=1}^{\infty}\left(1 - \left|E\exp\{i\tau\xi_m\}\right|^2\right)$$

We suppose that $\left|x_n\right| \le c$ Remark that as a rule the parameters of a specified allocation schemes can be chosen so that $x_n = 0$.

Theorem 3.1:
There exist constant C > 0 such that

$$\Delta_n = \sup_{-\infty<u<\infty} \left|\mathrm{P}_n(u) - \Phi(u)\right| \le C\left[\beta_{3,n} + \kappa_{3,n} + B_n^2\kappa_{3,n}\exp\left\{-M_n\left((4B_n\kappa_{3,n})^{-1}\right)\right\}\right.$$

$$+\frac{1}{\sqrt{M_n\left((4B_n\kappa_{3,n})^{-1}\right)}}\max\left(1,\frac{\min(B_n,\beta_{1,n})}{\sqrt{M_n\left((4B_n\kappa_{3,n})^{-1}\right)}}\right)\Bigg]$$

The application of Theorem 3.1 to aforementioned examples of GAS gives lower estimation of Δ_n .

IV. Applications:

4.1. Sample Scheme Without Replacements:

Let us consider the sample scheme without replacement from stratified population of size D_N as in Example 1, hence use that denotes; in particular d_m is the size of the m-th stratum, η_m is the number of elements of the m-th stratum appeared in the sample of size n; $\mathcal{L}(\xi_m) = Bi(d_m, p)$ m = 1,...,N. It is convenient to choose $p = n/D_N$ $q = 1 - p$ then $x_N = 0$ We allow d_m to be increased together with N. According to our best knowledge such case is considered for the first time.

Theorem 4.1:

There exist a constant C > 0 such that

$$\Delta_N \le C\left(\beta_{3,N} + \left(\frac{d}{nq}\right)^{1/2}\right)$$

where $d = \max_{1\le m\le N} d_m$

Let the elements of a stratum with index m are independent r.v.s $X_{m,1},\ldots,X_{m,d_m}$ having common distribution as of a r.v. Y_m , m = 1,.....N. Assume that $Y_1,\ldots,Y_N$ are independent r.v.s. Let $S_{n\,N}$ be a sample sum, i.e. sum of elements of the population that appeared in a sample. The r.v. $S_{n\,N}$ is type (1.2) with $f_m(0) = 0$, $f_m(j) = X_{m,1} + \ldots + X_{m,j}, j = 1,\ldots,N$ i.e.

$$S_{n,N} = \sum_{m=1}^{N}\left(\sum_{j=1}^{\eta_m} X_{m,j} \Uparrow \{\eta_m \ge 1\}\right)$$

Set $\gamma_N = \frac{1}{D_N}\sum_{m=1}^{N} d_m EY_m$, $\sigma_N^2 = p\sum_{m=1}^{N} d_m\left(\alpha_{2,0,m} - p\alpha_{1,0,m}^2\right)$

$$\rho_{j,N} = \max(1,(dp)^j)\frac{1}{\sigma_N^j}\sum_{m=1}^{N} E\left|Y_m - \gamma_N\right|^j$$

Theorem 4.2:

There exist a constant C >0 such that $\max_{-\infty<u<\infty}\left|\mathrm{P}_N(u) - \Phi(u)\right| \le C\left(\rho_{3,N} + \frac{\sqrt{d}}{\sqrt{nq}} + \frac{d^{3/2}}{\sqrt{n}}\right)$

Theorems 4.2 improves the results of Zhao, L. C., Wu,C. Q. and Wang Q.(2004).

4.2. Multinomial Allocation Scheme:

Consider the multinomial allocation scheme described in Example 2, hence use that denotes:

$\mathcal{L}(\xi_m) = \Pi(zp_m)$ P_m =0, m = 1,2 ; $p_1 + p_2 + \ldots = 1$ it is convenient to put z = n , then X_N = 0. Put $P(n) = \sum_{m=1}^{\infty} p_m^2$

Theorem 4.3:

There exist a constant C > 0 such that $\max_{-\infty<u<\infty} \left|\mathrm{P}_n(u) - \Phi(u)\right| \le C\left(\beta_{3,n} + \sqrt{\frac{1}{n} + \sum_{m=1}^{\infty} p_m^2}\right)$

4.3 Probabilistic Set-up Without Memory:

One of application of Theorem 4.3 is the following. Consider a probabilistic set-up without memory, input of which is sequence of independent identical distributed discrete random variables X_1.....X_N, on output we have a sequence of independent random variables Y_1,.....Y_N with two possible values 0 and 1 only.

Let $P\{X_j = x_m\} = p_m$ $p_1 + p_2 + \ldots = 1$ and $P\{Y_j = 1 / X_j = x_m\} = q_m$

$m \in \aleph$ Also, let η_{mj} be a frequency of the event $(X_l = x_m, Y_l = j)$ and $\eta_m = \eta_{m0} + \eta_{m1}$ a frequency of the event $(X_l = x_m)$ among the sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ We wish to test a simple hypothesis

$$H_0 : \{q_m = 1/2,\ m \in \aleph\} \tag{4.1}$$

versus to the sequence of alternatives

$$H_{1n} : \sum_{m=1}^{\infty} p_m^2 k_m^2 = K_n^2 > 0 \tag{4.2}$$

where $k_m = 2q_m - 1$ We consider a test based on statistic

$$X^2(n) = \sum_{m=1}^{\infty} (\eta_{m1} - \eta_{m0})^2$$

The large values of X^2 (n) rejects hypothesis H_0. Put $P(n) = \sum_{m=1}^{\infty} p_m^2$ From Theorem 4.3 we get:

Corollary 4.1. If as $n \to \infty$

$$n^2 P(n) \to \infty, \text{ and } \max_{m \in \aleph} p_m = o\left(\sqrt{P(n)}\right), \tag{4.3}$$

then for the X^2 (n) test of size α ε (0,1) we have:

(i) the critical point is $c,= n(1+u_{\alpha}\sqrt{2P(n)})$

(ii) the asymptotic power is $\Phi,(-u_{\alpha}+\frac{nK_n^2}{\sqrt{2P(n)}})$

where Φ (u) is standard normal distribution function, and $u_{\alpha}=\Phi^{-1}(1-\alpha)$ Thus this test detects the alternatives at the "distance" $\sqrt{2P(n)}/n$ from H_0.

Condition (4.3) imply that among all values of the input r.v. X_i must be enough many values having probabilities which comparable with $\max\limits_{m\in\aleph} p_m$

By Ivanov and Lapin (1983) and Ivanov (1986) has been considered the problem of testing H_0 (4.1) in the case when input random variables are simple, which corresponds to case when $p_m=1/N$ m = 1,2,.....N and $P_m = 0$ for all m > N. While $N\to\infty$as $n\to\infty$their result can't be used for the described above more general situation. Let us now consider the case of Ivanov and Lapin (1983) and Ivanov (1986). In this case q_m =q does not depend on *m*. Thus hypothesis H_0 is *q* =1/2 and the alternative H_{1n} is $k_m\equiv k=\Delta(n\alpha)^{-1/4}$, *m*=1,2,…, where α = n / N and $\Delta > 0$ is constant. The statistic X^2 *(n)* in this case has form $X^2(n,N)=\sum_{m=1}^{N}(\eta_{m1}-\eta_{m0})^2$

Corollary 4.2:
If $n\alpha\to\infty$ and $\alpha=o(n^{1/3})$ then the statistic $X^2(n)$ has an asymptotical normal distribution with expectation $A(n,N)=n(1+\alpha k^2)$ and variance $\sigma^2(n,N)=2n\alpha(1+2\alpha k^2(1-k^2))$.

Corollary 4.3:

If $n\alpha\to\infty$ and $\alpha=o(n^{1/3})$. then the critical region of the $X^2(n,N)$ test of size ω ε (0,1) is $\left\{X^2(n,N)>n+u_{\omega}\sqrt{2n\alpha}\right\}$ and asymptotical power is .$\Phi\left(-u_{\omega}+\frac{\Delta^2}{\sqrt{2}}\right)$

These results are complement to those of Ivanov and Lapin (1983), where additionally has been assumed that $\alpha\geq c>0.$

REFERENCES

Bhattachatya, R.N. and R. Ranga Rao, 1976. Normalapproximation and asymptotic expansions. John Wiley& Sons, New-York.

Ivanov VA., 1986. On generalization of the decomposablestatistics. Theory Probabl. Appl., 21: 786-790.

Ivanov V.A. and C.A. Lapin, 1983 Asymptotical normalityof the randomized decomposable statistics in the multinomial scheme. Math. Notes, 39: 745-756.

Jonson, N.L., Kotz, S. Urn 1977. models and theirapplications. Wiley, New York.

Kotz, S. and N. Balakrishan, 1997. Advances in urn modelsduring the past two decades. In Advances in Combinatorial Methods and Appl. to Probabl. and Statist., pp: 203-257. Birkhauser, Boston. MA.

Mirakhmedov Sh., 1992. Randomized decomposable statistics in the scheme of independent allocating particles into boxes., 2: 91-108.

Mirakhmedov, Sh., 1996. Limit theorems on decomposable statistics in a generalized allocation schemes. Discrete Math. Appl., 6: 379-404.

Mirakhmedov, Sh., 2007. Asymptotic normality associated with generalized occupancy problem. Statist. & Probab. Letters, 77: 1549-1558.

Mirakhmedov, S.S. and Sh. Mirakhmedov, 2009. On asymptotic expansion in the random allocation of particles by sets. J. Theoretical Probab. 22: DOI-10.1007/s10959-009-0225-7.

Zhao, L.C., C.Q. Wu, and Q. Wang, 2004. Berry-Esseen bound for a sample sum from a finite set of independent random variables. J. Theoretical Probab. 17: 557-572.